

LLaDa: Fine-tuned LLM for Data Science

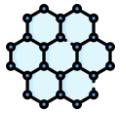
2021320301 최성아

2021320303 정지원

2021320307 김은진

2021320322 윤민서





GRAPHENE: GRaph-based Analysis and Prediction

Harnessing Efficient Natural language Elucidation in Data Science



2021320301 최성아

2021320303 정지원

2021320307 김은진

2021320322 윤민서

MATRIX: Massive and Advanced Talking-model Reas oning Interview eXpertise in data science

2021320301 최성아

2021320303 정지원

2021320307 김은진

2021320322 윤민서

Agenda

- Introduction
 - Goal (해결하고자 하는 과제)
 - Data collection & preprocessing (데이터 수집 방법)
 - Connection (본 과목과의 연관성)
- Methodology
- Experiment
- Conclusion

Introduction

Lost Large Money! 💰

12월 11일 월요일

-8,901원
내 카카오뱅크계좌 → (US)OPENAI
+14158799686 USA

12월 9일 토요일

-8,874원
내 카카오뱅크계좌 → (US)OPENAI
+14158799686 USA

11월 21일 화요일

-7,263원
내 카카오뱅크계좌 → (US)OPENAI
+14158799686 USA

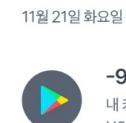
11월 19일 일요일

-7,287원
내 카카오뱅크계좌 → (US)OPENAI
+14158799686 USA



Invoices

Dec 2023	Paid	\$6.60	>
Dec 2023	Paid	\$6.60	>
Nov 2023	Paid	\$5.50	>
Nov 2023	Paid	\$5.50	>



-988원
내 카카오뱅크계좌 → (US)GOOGLE *Google Storage 650-253-0000 USA



+329원
GOOGLE *Colab → 내 카카오뱅크계좌



-66,940원
내 카카오뱅크계좌 → (US)GOOGLE *Colab 650-253-0000 USA

11월 30일 목요일

-13,253원
내 KB국민계좌 → 마스타해외승인출금

11월 24일 금요일

-66,621원
내 KB국민계좌 → 마스타해외승인출금

-67,451원
내 카카오뱅크계좌 → (US)GOOGLE *Colab cc@google.com USA

7일 목요일



-13,421원
내 카카오뱅크계좌 → (US)GOOGLE *Colab cc@google.com USA

6일 수요일



-13,227원
내 카카오뱅크계좌 → (US)GOOGLE *Colab cc@google.com USA

3일 일요일

Introduction

- Background
 - 정보의 과부하는 데이터 사이언스 분야의 주요한 문제임
 - 데이터 사이언스 분야는 컴퓨터과학, 통계학, 수학적 지식을 기반으로 머신러닝 및 딥러닝 지식을 이해할 수 있어야 함
 - 오늘날 많은 지식을 검색 엔진을 통해 얻을 수 있지만, 데이터 사이언스 분야만의 지식을 얻기는 어려움
 - 검색 엔진과 달리, 언어 모델은 데이터 사이언스 분야의 지식을 저장하고, 활용하고,
추론할 수 있는 능력이 있음

Introduction

- Goal
 - Field-specific fine-tuning을 통해 언어 모델이 데이터 사이언스 지식을 쌓을 수 있도록 학습함
 - 사용자가 언어 모델과의 대화 또는 언어 모델의 추론 결과를 통해 데이터 사이언스 지식을 습득할 수 있도록 함
 - 다양한 fine-tuning 방법론들을 이 프로젝트에 적용하고 평가함

Introduction

- Data Collection
 - Less Is More for Alignment 논문에서, 적지만 양질의 데이터를 활용해서 fine-tuning 하면 더 좋은 결과를 얻는다는 것을 보여줌
 - 인터넷에서 단순히 크롤링하는 방식이 아닌, 직접 양질의 데이터를 모으기로 결정
 - 데이터 과학 분야에서의 면접 질문과 답변으로 이루어진 데이터를 수집
 - Question Answering Task
 - 면접 질문은 데이터 사이언스 지식의 핵심 내용이 들어있기 때문임
 - 1. 사용자가 대화 기반으로 데이터 사이언스 지식을 습득할 수 있도록

Introduction

• Data References

Published as a conference paper at ICLR 2021

MEASURING MASSIVE MULTITASK LANGUAGE UNDERSTANDING

Dan Hendrycks
UC Berkeley

Collin Burns
Columbia University

Steven Basart
UChicago

Andy Zou
UC Berkeley

Mantas Mazeika
UIUC

Dawn Song
UC Berkeley

Jacob Steinhardt
UC Berkeley

Deep Learning Interviews is home to **hundreds** of fully-solved problems, from a **wide range of key topics in AI**. It is designed to both rehearse **interview or exam-specific topics** and provide machine learning **M.Sc./Ph.D. students**, and those awaiting an interview a well-organized overview of the field. The problems it poses are tough enough to **cut your teeth on** and to dramatically improve your skills—but they're framed within **thought-provoking** questions and engaging stories.

That is what makes the volume so specifically valuable to students and job seekers: it provides them with the ability to speak confidently and quickly on any relevant topic, to answer technical questions clearly and correctly, and to fully understand the purpose and meaning of interview questions and answers. These are powerful, indispensable advantages to have **when walking into the interview room**.

The book's contents is a large inventory of numerous topics relevant to DL job interviews and graduate-level exams. That places this work at the forefront of the growing trend in science to teach a core set of practical mathematical and computational skills. It is widely accepted that the training of every computer scientist must include the fundamental theorems of ML, and AI appears in the curriculum of nearly every university. **This volume is designed as an excellent reference for graduates of such programs.**

Shlomo Kashani, Author. Amir Irvy, Chief Editor.



www.interviews.ai

DEEP LEARNING INTERVIEWS

- Logistic Regression
- Information Theory
- Calculus
- Algorithmic Differentiation
- Bayesian Deep Learning
- Probabilistic Programming
- Ensemble Learning
- CNN Feature Extraction
- Deep Learning: Expanded Chapter



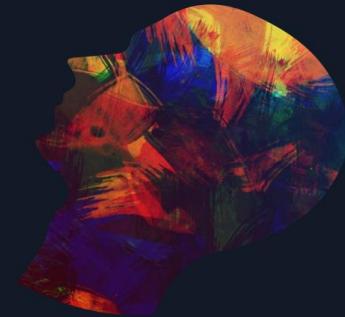
DEEP LEARNING INTERVIEWS

SHLOMO KASHANI

DEEP LEARNING INTERVIEWS

—REAL-WORLD DEEP LEARNING INTERVIEW
PROBLEMS & SOLUTIONS—

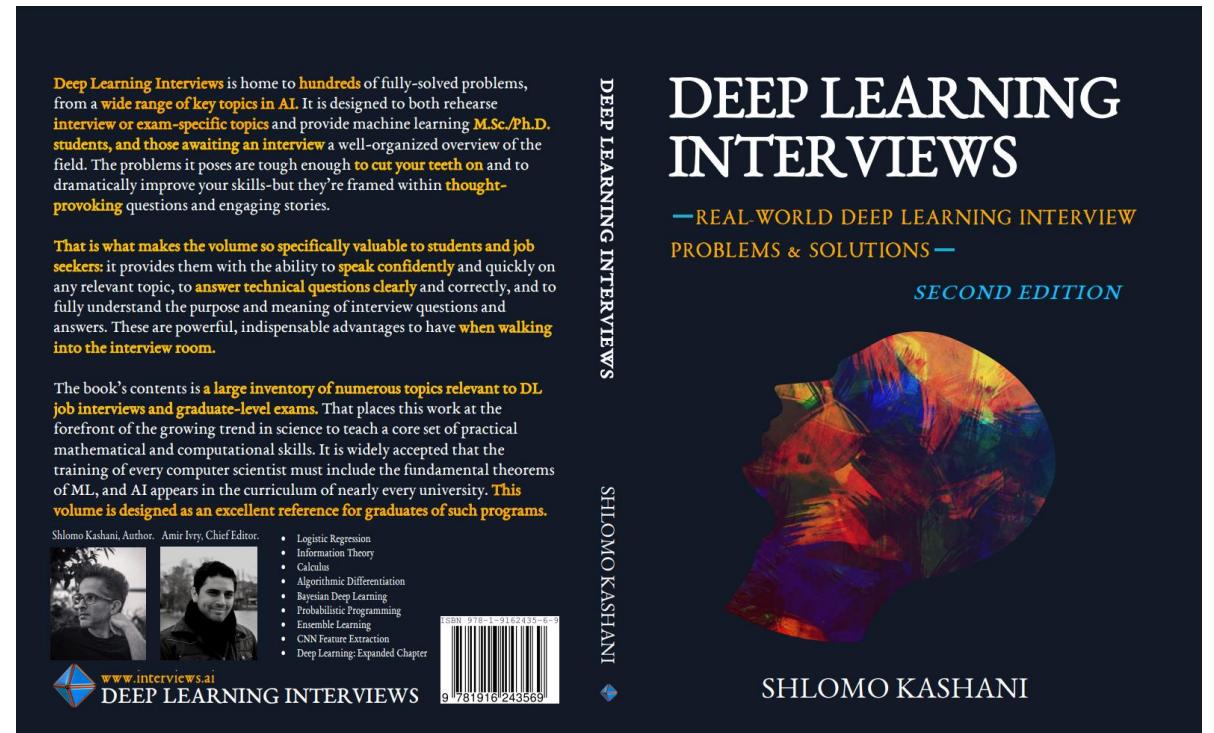
SECOND EDITION



SHLOMO KASHANI

Introduction

- Data References
 - Interview QA
 - 500 Computer Science
 - 500 Statistics
 - 500 AI
 - 직접 수집
 - Hand-craft preprocessing



Introduction

- Data References

Published as a conference paper at ICLR 2021

MEASURING MASSIVE MULTITASK LANGUAGE UNDERSTANDING

Dan Hendrycks
UC Berkeley

Collin Burns
Columbia University

Steven Basart
UChicago

Andy Zou
UC Berkeley

Mantas Mazeika
UIUC

Dawn Song
UC Berkeley

Jacob Steinhardt
UC Berkeley

- MMLU Dataset

- 과학, 역사, 문학, 전문직 등 57개의 서로 다른 주제
- 구성: MCQA / 4지선다
- 모델이 단순한 언어 패턴 인식을 넘어 실제 지식과 추론 능력을 갖추고 있는지를 확인

Introduction

- Data Preprocessing – Interview QA
 - Interview question, Conceptual question을 수집
 - 500 Computer Science – 컴퓨터구조, 운영체제, 컴퓨터네트워크 등
 - 500 Statistics – 확률론, 수리통계, 회귀분석 등
 - 500 AI – 머신러닝 이론 및 기법, 기초 딥러닝, 컴퓨터비전, 자연어처리 등
 - 수식 표현에서의 그리스 문자를 직접 처리
 - 데이터 표현에서의 쌍따옴표 등을 직접 처리 (JSONL 파일 형식)
 - 언어 (내지는 코드, 수식) 형식으로 표현하기 부적절한 데이터 제거
 - 총 1481개의 데이터를 사용, 1000 training set, 481 testing set

Introduction

- Data Preprocessing – MMLU Dataset
 - High School Computer Science, College Computer Science
 - High School Mathematics, College Mathematics
 - High School Statistics, Machine Learning
- 6개 분야의 dev, test, val 데이터를 모아서 다시 train-test split
 - 1022개의 데이터 – 각각 817, 205개로 split

Introduction

- Connection
 - 9주차 정보 추출
 - 지식 그래프
 - 12주차 자연어 생성, 13주차 대화 시스템
 - LLaMA을 활용한 instruction fine-tuning

Methodology

- Task 1: 직접 수집한 데이터셋으로 주관식 문제 평가
 - Generative Question Answering Task
- Task 2: MMLU 데이터셋으로 객관식 문제 평가
 - Answer Selection Task

Methodology: Task 1

- Task 1: 직접 수집한 데이터셋으로 주관식 문제 평가
- 질문이 주어지면 이에 상응하는 대답을 생성하는 Task

Example Question & Answer

Question. What do you understand by Natural Language Processing?

Answer: Natural Language Processing is a field of computer science that deals with communication between computer systems and humans. It is a technique used in Artificial

Intelligence and Machine Learning. It is used to create automated software that helps understand human-spoken languages to extract useful information from the data. Techniques in NLP allow

computer systems to process and interpret data in the form of natural languages.

Methodology: Task 1

- Task 1: 직접 수집한 데이터셋으로 주관식 문제 평가
- 시도한 방법론
 - Llama2 7B (Baseline)
 - Fine-tuned Llama2 7B
 - Llama2 7B + Instruction
 - Instruction-tuned Llama2 7B
 - Instruction-tuned Llama2 7B + Simplified GoT

Methodology: Task 1

- Task 1 – Fine-tuning (mid)
 - 단순한 설명과 질문/답변을 프롬프트로 사용해서 학습

Prompt

You are applying for a job related to AI, and you can expect to encounter problems related to statistics, computer science, and artificial intelligence.

Question:

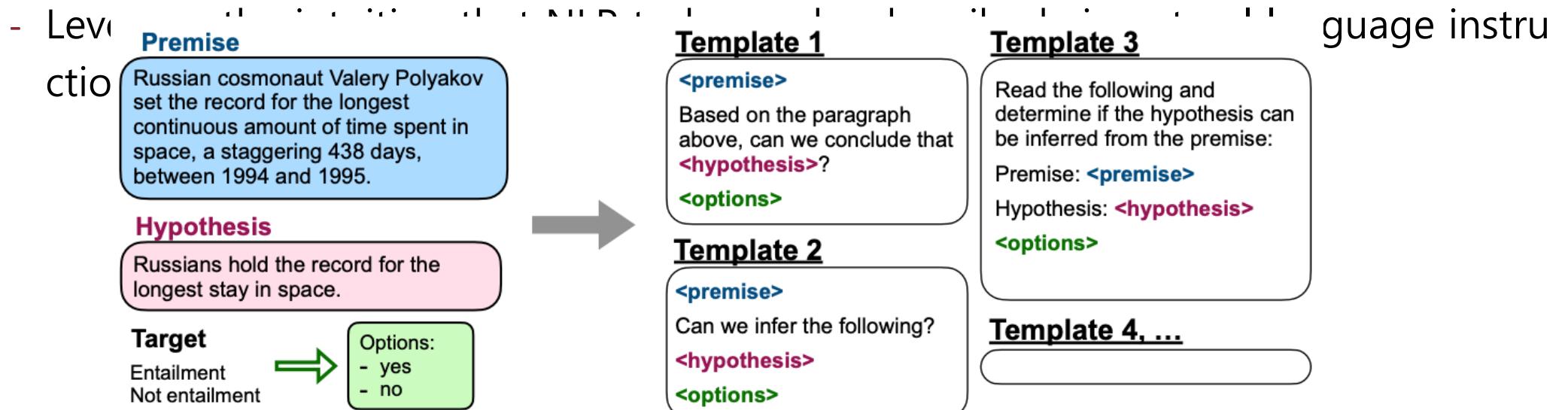
{Question}

Answer and Explanation:

{Answer}

Methodology: Task 1

- Task 1 – Instruction Tuning
 - Reference: Finetuned Language Models Are Zero-Shot Learners (ICLR 2022)



Methodology: Task 1

- Task 1 – Instruction Tuning
 - Llama2 의 Instruction, System 토큰 사용, Alpaca instruction 문구 사용

Instruction Prompt

Below is an instruction that describes a task. Write a response that appropriately completes the request.

[INST] <<SYS>>

You are a job applicant with expertise in Mathematics, Statistics, Computer Science, and Machine Learning.
Please write the answer of the following interview question with a clear and comprehensive explanation to demonstrate depth of knowledge.

Please answer only the given one question without any additional information.

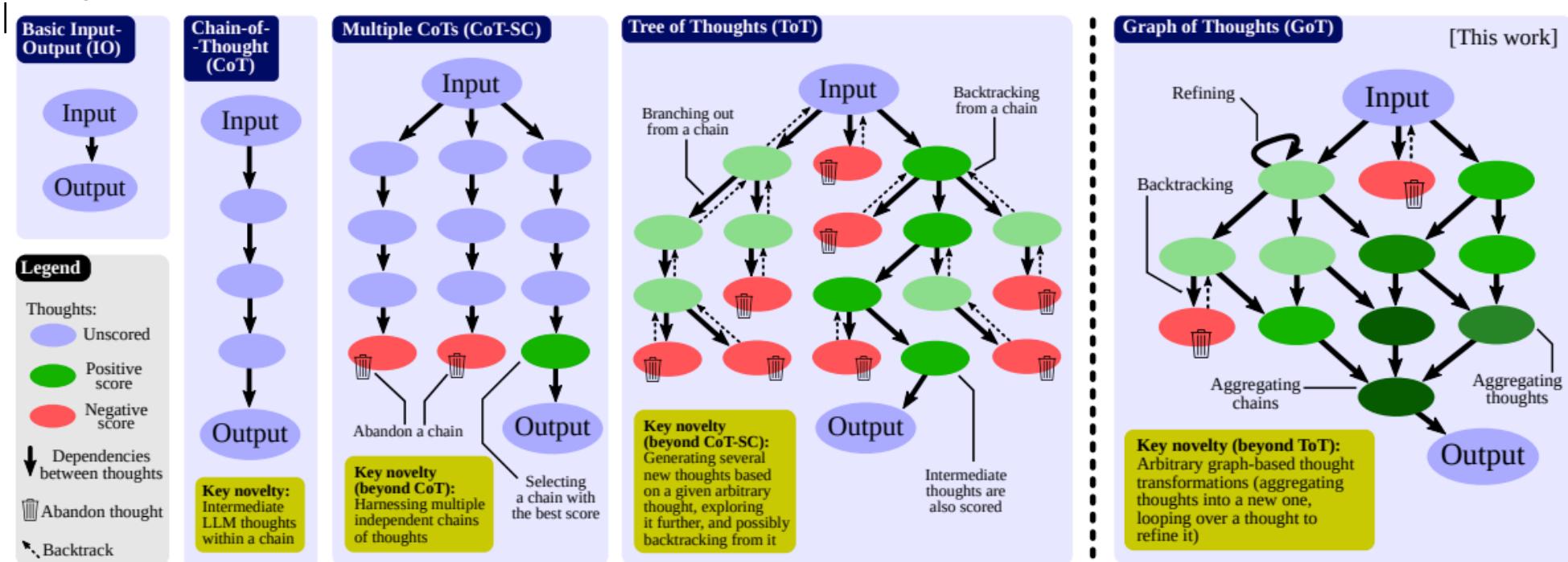
<</SYS>>

Question: {{Question}} **[/INST]**

Answer: {{Answer}}

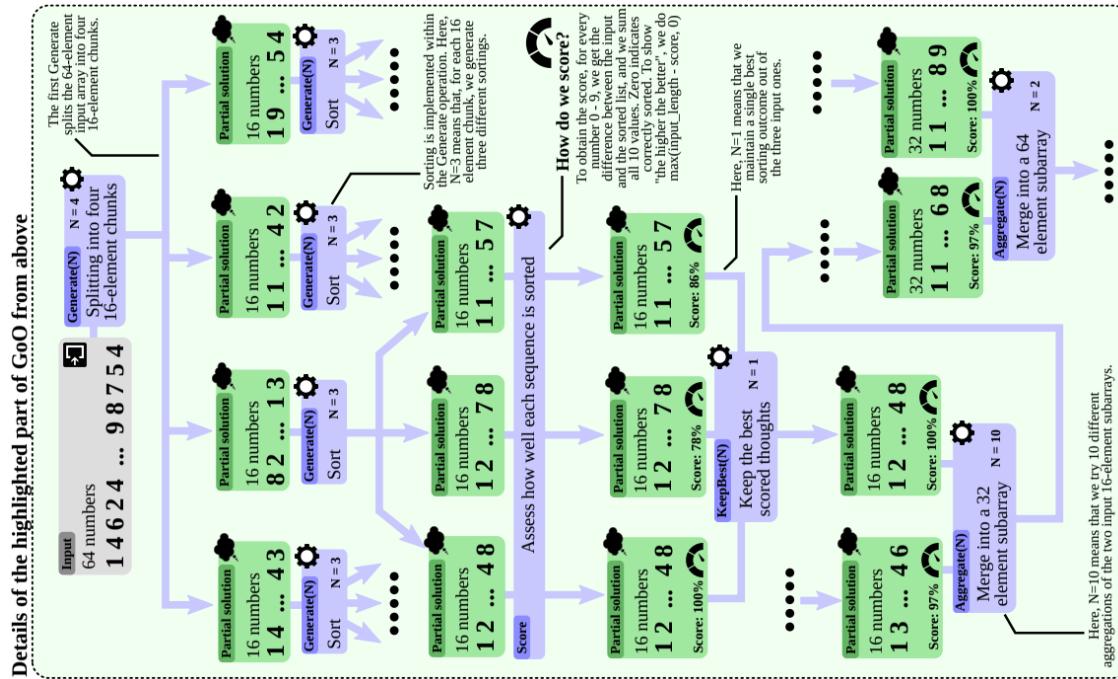
Methodology: Task 1

- Task 1 – Graph-of-Thoughts
 - Reference: Graph of Thoughts: Solving Elaborate Problems with Large Language



Methodology: Task 1

- Task 1 – Graph-of-Thoughts
 - Reference: Graph of Thoughts: Solving Elaborate Problems with Large Language Models



Methodology: Task 1

- Task 1 – Simplified Tree-of-Thoughts

In ToT Prompt ~~ent experts are answering this question.~~

All experts will write down 1 step of their thinking, then share it with the group.

Then all experts will go on to the next step, etc.

If any expert realises they're wrong at any point then they leave.

The question is...

Bob is in the living room.

He walks to the kitchen, carrying a cup.

He puts a ball in the cup and carries the cup to the bedroom.

He turns the cup upside down, then walks to the garden.

He puts the cup down in the garden, then walks to the garage.

Where is the ball?

Answer

Expert 1: The ball is in the living room.

Expert 2: The ball is in the kitchen.

Expert 3: The ball is in the bedroom.

Expert 1: Bob carries the cup to the bedroom, so the ball must be in the cup.

Expert 2: Oh, I see my mistake. Yes, the ball is in the cup.

Expert 3: Agreed, the ball is in the cup in the bedroom.

Expert 1: Next, Bob turns the cup upside down in the bedroom.

Expert 2: After that, Bob walks to the garden and puts the cup down.

Expert 3: Therefore, the ball must have fallen out of the cup when Bob turned it upside down.

Expert 1: Oh, you're right. I made a mistake. The ball is in the bedroom, not in the cup.

Expert 2: Agreed, the ball is in the bedroom.

Expert 3: Bob then walks to the garage, so the ball remains in the bedroom. It is not in the

Expert 1: Absolutely, the ball is still in the bedroom.

Expert 2: Yes, the ball hasn't moved from the bedroom.

Expert 3: Therefore, the ball is in the bedroom, not in the garden or the garage.

All three experts agree that the ball is in the bedroom.

Methodology: Task 1

- Task 1 – Simplified Graph-of-Thoughts
 1. 주어진 질문에 대해 핵심 키워드를 선정
 2. 핵심 키워드들에 대해 설명하도록 요청
 3. 처음 주어진 질문 + 키워드 + 설명을 기반으로 질문에 답변하도록 요청

Methodology: Task 1

- Task 1 – Simplified Graph-of-Thoughts

- 주어진 질문에 대해 핵심 키워드를 선정

GoT Prompt 1

Below is an instruction that describes a task. Write a response that appropriately completes the request.

[INST] <<SYS>>

You have expertise in Mathematics, Statistics, Computer Science, and Machine Learning.

Please write the keywords of the following interview question to use your knowledge.

<</SYS>>

Example:

Question: What are the applications of Deep Learning?

Keywords of question:

Deep Learning, Computer vision, Natural language processing (NLP), Reinforcement learning

Referencing the example above, find the key keywords in the below question.

Question: {{Question}} [/INST]

Keywords of question:

Methodology: Task 1

- Task 1 – Simplified Graph-of-Thoughts

2. 핵심 키워드들에 대해 설명하도록 요청

GoT Prompt 2

Below is an instruction that describes a task. Write a response that appropriately completes the request.

[INST] <<SYS>>

You have expertise in Mathematics, Statistics, Computer Science, and Machine Learning.

Please provide a simple but clear explanation for each of the keywords below to demonstrate depth of knowledge.

<</SYS>>

Keywords: {{Keywords}} [/INST]

Explanation of keywords:

Methodology: Task 1

- Task 1 – Simplified Graph-of-Thoughts

3. 처음 주어진 질문 + 키워드 + 설명을 기반으로 질문에 답변하도록 요청

GoT Prompt 3

Below is an instruction that describes a task. Write a response that appropriately completes the request.

[INST] <<SYS>>

You are a job applicant with expertise in Mathematics, Statistics, Computer Science, and Machine Learning.

Please write the answer of the following interview question with a clear and comprehensive explanation to demonstrate depth of knowledge.

Please answer only the given one question without any additional information.

Question: {{Question}}

Keywords of question: {{Keywords}}

Explanations of keywords:

 {{Explanations}}

<</SYS>>

Only refer to keywords or explanations when necessary; focus on providing an accurate answer to the given question.

[/INST]

Answer:

Methodology: Task 1

- Task 1: 주관식 답 채점 방식
 - NIST
 - METEOR
 - G-Eval

Methodology: Task 1

- Task 1 채점 방식: NIST-MT
 - 기계 번역 metric
 - BLEU에 기반 (n-gram precision)
 - n-gram precision의 산술 평균
 - 개선된 짧은 문장 길이에 대한 페널티 (brevity penalty)

Methodology: Task 1

- Task 1 채점 방식: METEOR
 - 기계 번역 metric
 - BLEU 단점 보완
 - 동일한 단어를 찾고 주변 유의어를 확인
 - recall 과 precision 의 조화 평균 (recall 에 더 가중됨)

Methodology: Task 1

- Task 1 채점 방식: G-Eval
 - 가장 성능이 좋다고 알려진 LLM인 GPT-4가 직접 평가
 - 최근 LLM들이 많이 쓰고 있는 metric
 - SOTA evaluator들 보다 성능이 좋음
 - 사람이 평가한 것과 비슷한 모습을 보여줌

Methodology: Task 1

- Task 1 채점 방식: G-Eval
 - 프롬프트에 질문과 답변을 넣은 뒤 GPT-4에게 점수(1~5)로 채점하라고 요청
 - Score Function in G-Eval
$$\text{score} = \sum_{i=1}^n p(s_i) \times s_i$$
 - Score를 여러개 뽑은 뒤 평균을 내는 방식
 - Temperature=2, top_p=1로 설정 후 10개를 뽑아 점수를 평균냄

G-Eval Prompt

You will be given one question and one answer. Your task is to rate the answer on one metric. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:

Factual Consistency (1-5): Does the answer contain untruthful or misleading facts that are not supported by the question?

Evaluation Method:

Rate the factual consistency of the answer on a scale of 1 to 5, where 1 indicates a low level and 5 indicates a high level of consistency with the question.

- 1: Poor. The answer contains multiple inaccuracies or misleading details.
- 3: Fair. Some elements of the answer lack factual support or contain minor inconsistencies, but the majority aligns with the question's context.
- 5: Good. The answer is factually consistent, presenting information supported by the question without any misleading details.

Question:

`{{Question}}`

Answer:

`{{Answer}}`

Evaluation Form (scores ONLY):

- Factual Consistency:

Methodology: Task 2

- Task 2: MMLU 데이터셋으로 객관식 문제 평가
 - 답은 A, B, C, D 중 하나
 - LLM이 최대 토큰 수에 막혀 최종 답변까지 생성하기 전에 출력이 잘렸거나, 답변을 생성하지 않고 다른 말을 한 경우에는 E를 선택한 경우로 간주
 - 즉, 평가는 5지 선다 문제인 것으로 생각하고 진행함
 - LM + Reasoning layer 모델의 경우에도 5개 중에 1개를 선택하도록 설계

Methodology: Task 2

- Task 2: MMLU 데이터셋으로 객관식 문제 평가
 - 답은 A, B, C, D 중 하나

Example Question & Answer

"Question": "A correlation of 0.6 indicates that the percentage of variation in y that is explained by the variation in x is how many times the percentage indicated by a correlation of 0.3?"

"A": "2", "B": "3", "C": "4", "D": "6"

Methodology: Task 2

- Task 2: MMLU 데이터셋으로 객관식 문제 평가
 - 답은 A, B, C, D 중 하나

Example Question & Answer

"Question": "A correlation of 0.6 indicates that the percentage of variation in y that is explained by the variation in x is how many times the percentage indicated by a correlation of 0.3?"

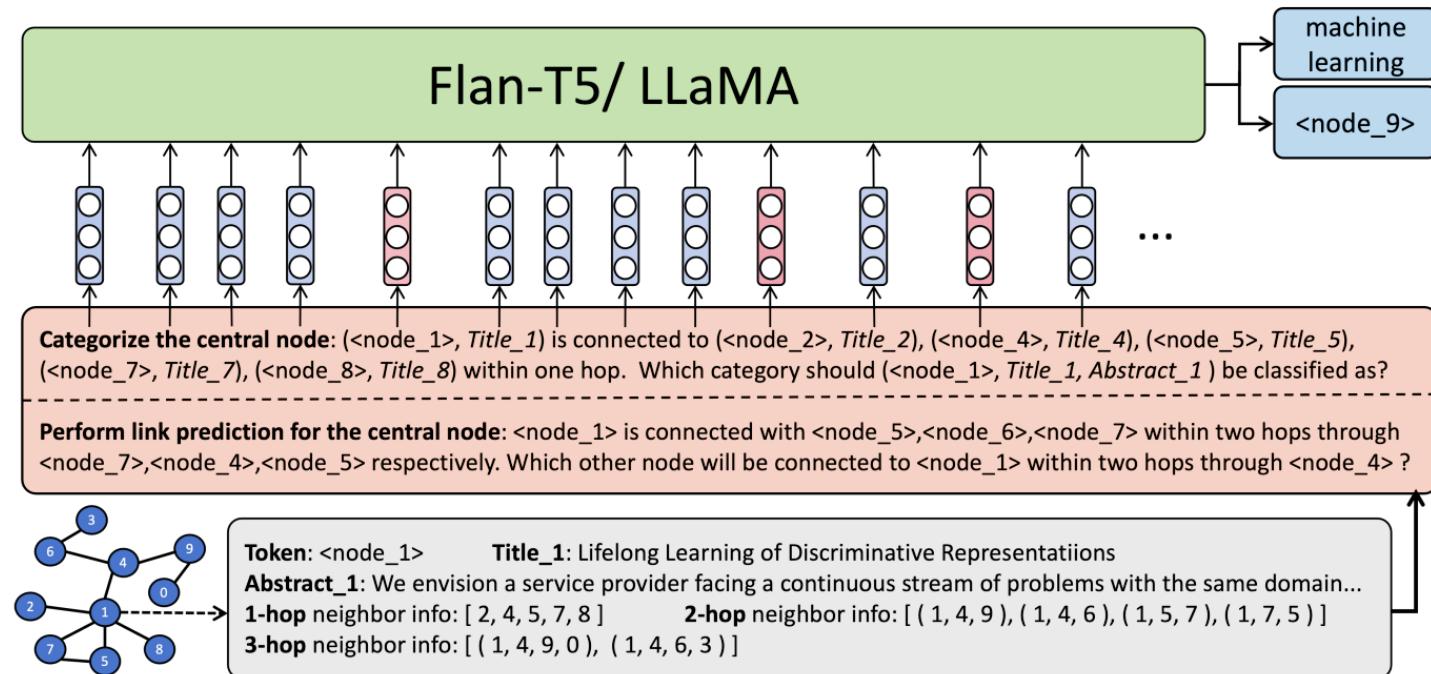
"A": "2", "B": "3", "C": "4", "D": "6", "Answer": "C"

Methodology: Task 2

- Task 2: MMLU 데이터셋으로 객관식 문제 평가
- 시도한 방법론
 - Llama2 7B (Baseline)
 - Fine-tuned Llama2 7B
 - Fine-tuned Llama2 7B + CoT (Why? → Experiment)
 - Llama2 7B + Knowledge Graph Instruction
 - RoBERTa + Knowledge Graph + Graph Neural Network
 - GPT-3.5, GPT-4

Methodology: Task 2

- Task 2 - Knowledge Graph Instruction
 - Reference: InstructGLM (arXiv 2023, Under Review)



Methodology: Task 2

- Task 2 - Knowledge Graph Instruction

- Prompt 정보

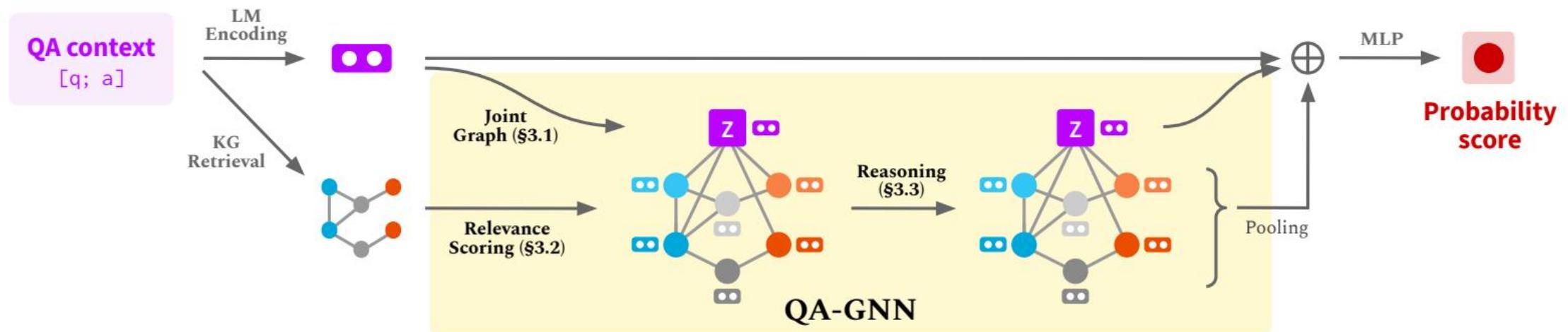
1. 문제에서 가장 중요한 키워드를 추출
2. 지식 그래프에서 추출한 키워드의 이웃의 정보까지 제공

- 중요한 키워드의 기준 → TF-IDF

- 문제에 등장하는 단어 중 TF-IDF가 높은 top-k개 단어 중 n-hop의 단어를 추출
- 추출된 단어들 중에서도 TF-IDF가 높은 top-k개 단어를 사용
- top-k개 단어 중 정수, 실수, sqrt, frac, 숫자 * 미지수 ($2x$, $3a$, ...) 형태는 제외

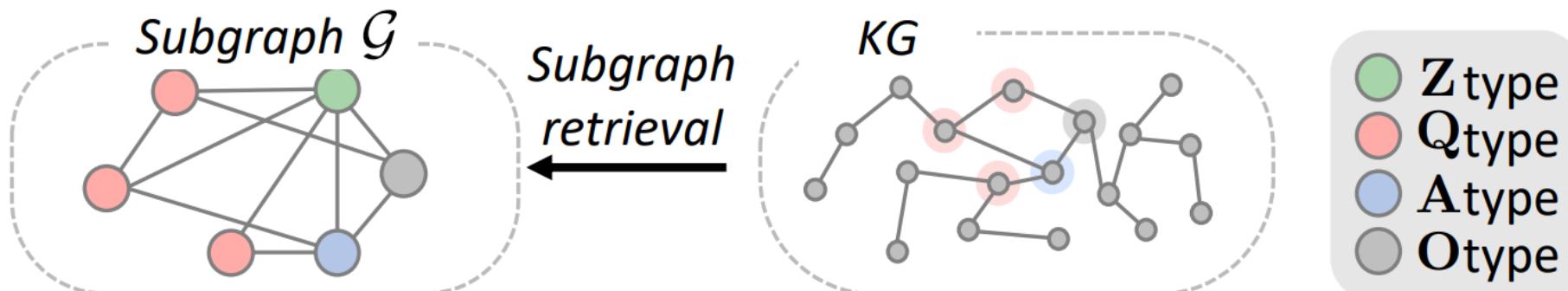
Methodology: Task 2

- Task 2 - RoBERTa + Knowledge Graph + Graph Neural Network
 - Reference: QA-GNN (NAACL 2021)



Methodology: Task 2

- Task 2 - RoBERTa + Knowledge Graph + Graph Neural Network
 - 공정한 평가를 위해서 E: ‘None of the above’ 선택지 추가
 - 계산 문제에서 등장하는 숫자를 지식 그래프에 모두 넣는 것은 불가능
 - 특수 토큰 추가: [1]
 - 숫자, 기호로만 구성된 문제일 경우에 Q 집합, A 집합이 공집합이 되는 현상을 방지



Experiment: Task 1

- Task 1: 직접 수집한 데이터셋으로 주관식 문제 평가

Model	GPT-4 Score	METEOR	NIST-MT
Llama2 7B (Baseline)	3.922	0.2620	1.1904
Fine-tuned Llama2 7B	3.664	<u>0.2640</u>	1.4353
Llama2 7B + Instruction	<u>3.923</u>	0.2548	1.0095
Instruction-tuned Llama2 7B	4.142	0.2675	1.1812
Llama2 7B + GoT	3.288	0.2383	<u>1.3045</u>

Experiment: Task 1

- Task 1 – Setting Colab Pro+
- Llama2 7B
 - 4bit quantization, 16bit floating point, LoRA, PEFT
 - LIMA와 동일한 hyperparameter
 - Training: A100 40GB
 - Generation: T4

Experiment: Task 1

- Llama2 7B (Baseline)
 - GPT-4 Score: 3.922, METEOR: 0.2620, NIST-MT: 1.1904
- Fine-tuned Llama2 7B
 - GPT-4 Score: 3.664, METEOR: 0.2640, NIST-MT: **1.4353**
 - 단순한 프롬프트를 사용하였기에, 주어진 질문에 제대로 답을 하지 않고 새로운 질문과 답변을 계속 만들어나가는 문제가 있음
 - Instruction prompt에서 해결됨

Experiment: Task 1

- Llama2 7B + Instruction
 - GPT-4 Score: 3.923, METEOR: 0.2548, NIST-MT: 1.0095
 - GPT-4 Score 도 크게 변하지 않았고, METEOR와 NIST metric 이 감소함
 - 단순히 Instruction prompt 만으로는 오히려 안 좋은 성능을 낼 수 있음을 보여줌
- Instruction-tuned Llama2 7B
 - GPT-4 Score: **4.142**, METEOR: **0.2675**, NIST-MT: 1.1812
 - GPT-4 Score 및 METEOR metric 이 가장 높음
 - Instruction tuning 이 QA 태스크에서 매우 효과적임을 보여줌

Experiment: Task 1

- Instruction-tuned Llama2 7B + Simplified GoT
 - GPT-4 Score: 3.288, METEOR: 0.2383, NIST-MT: 1.3045
 - 모든 metric 에서 감소함
 - keyword + explanation 과정에서 오히려 질문의 의도를 흐리는 문제가 발생함

Experiment: Task 2

- Task 2: MMLU 데이터셋으로 객관식 문제 평가

모델	Accuracy	답변 생성 비율
Llama2 7B (Baseline)	7.32	21.46
Llama2 7B + Fine-Tuning	18.05	67.80
Llama2 7B + CoT	28.29	92.19
Llama2 7B + KG Instruct	28.78	94.15
RoBERTa + KG + GNN (QA-GNN)	32.68	100.00
GPT-3.5	36.00	100.00
GPT-4	65.37	100.00

Experiment: Task 2

- Task 2 - Setting Colab Pro+
- Llama2 7B
 - 4bit quantization, 16bit floating point, LoRA, PEFT
 - Training: A100 40GB
 - KG Construction, Generation: T4
- RoBERTa + GNN
 - KG Construction: A100 40GB
 - Training, Reasoning: T4

Experiment: Task 2

- Task 2 - Llama2 7B (Baseline)
 - Accuracy: 7.32%, 답변 생성 비율: 21.46%
 - 답변 생성 비율: 생성된 텍스트 중 선지의 정보가 들어있다는 것을 알 수 있는 것의 비율

Used Prompt

You are a job applicant in an interview, tackling questions that fall into one of the following categories: Mathematics, Statistics, Computer Science, or Machine Learning. For each question, you are presented with four options: A, B, C, and D. You must choose only one correct answer from these options, without any additional explanation or reasoning.

Question: {eval_entry['Question']}

A: {eval_entry["A"]}
B: {eval_entry["B"]}
C: {eval_entry["C"]}
D: {eval_entry["D"]}

Answer:

Experiment: Task 2

- Task 2 - Fine-tuned Llama2 7B
 - Accuracy: 18.05%, 답변 생성 비율: 67.80%
 - Training: LIMA와 동일한 hyperparameter
 - Evaluation: huggingface generate 메소드의 hyperparameter 수정 (보고서)
 - Baseline과 동일한 prompt를 사용하면 주어진 문제에 답을 하지 않고 새로운 질문과 보기를 계속 반복해서 만들어내는 경향이 있음 (무의미)
 - 이미 fine-tuning한 모델이지만 few-shot learning을 하는 것과 유사하게 prompt에 예시를 하나 추가하면 어떨까?
 - 소규모 LLM의 경우에는 fine-tuning, few-shot prompt를 동시에 사용하는 것이 효과적인 것으로 관찰

Used Prompt

You are a job applicant in an interview, tackling questions that fall into one of the following categories: Mathematics, Statistics, Computer Science, or Machine Learning.

For each question, you are presented with four options: A, B, C, and D.

Your sole task is to select one correct answer from these options for the provided questions, without creating new questions or providing additional explanations.

Example Question: In machine learning, what does the term 'gradient descent' refer to?

- A: A method for dividing data into clusters
- B: An algorithm for optimizing a function
- C: A technique for classifying data
- D: A type of neural network architecture

Answer: B

Question: {eval_entry['Question']}

- A: {eval_entry["A"]}
- B: {eval_entry["B"]}
- C: {eval_entry["C"]}
- D: {eval_entry["D"]}

Answer:

Experiment: Task 2

- Task 2 - Fine-tuned Llama2 7B + CoT
 - Accuracy: 28.29%, 답변 생성 비율: 92.19%
 - Evaluation: huggingface generate 메소드의 hyperparameter 수정 (보고서)
 - 왜 fine-tuning과 CoT를 같이 사용?
 - 앞 실험에서 관찰한 결과에 기인

Used Prompt

You are a job applicant in an interview, tackling questions that fall into one of the following categories: Mathematics, Statistics, Computer Science, or Machine Learning.

For each question, you are presented with four options: A, B, C, and D.

Your sole task is to select one correct answer from these options for the provided questions, without creating new questions or providing additional explanations.

Example Question: In machine learning, what does the term 'gradient descent' refer to?

- A: A method for dividing data into clusters
- B: An algorithm for optimizing a function
- C: A technique for classifying data
- D: A type of neural network architecture

'Gradient descent' is a term often used in optimization problems.

It involves iteratively moving towards the minimum of a function.

This process is not specific to dividing data, classifying data, or a neural network architecture.

Therefore, it aligns best with the concept of an algorithm for optimizing a function.

Answer: B

Question: {eval_entry['Question']}

- A: {eval_entry["A"]}
- B: {eval_entry["B"]}
- C: {eval_entry["C"]}
- D: {eval_entry["D"]}

Let's think step-by-step. # 원래는 one-shot 이후에는 step-by-step prompt를 넣지 않지만, 실험 결과 본 prompt의 성능이 더 좋다는 것이 확인됨.

Experiment: Task 2

- Task 2 - Llama2 7B + Knowledge Graph Instruct
 - Accuracy: 28.78%, 답변 생성 비율: 94.15%
 - TF-IDF가 높은 top-k개 단어의 n-hop의 단어 중 top-k개를 사용
 - $k_1 = 20, n = 3, k_2 = 10$
 - k_1 이나 n 이 지나치게 크면 특정한 단어들만 키워드로 선택되는 현상 발생
 - k_2 가 지나치게 작으면 모델에게 주어지는 정보가 적어짐

Used Prompt

You are a job applicant in an interview, tackling questions that fall into one of the following categories: Mathematics, Statistics, Computer Science, or Machine Learning.

For each question, you are presented with four options: A, B, C, and D.

Your sole task is to select one correct answer from these options for the provided questions, without creating new questions or providing additional explanations.

Additionally, use the given keywords to provide the explanation behind your answer selection.

Question: {eval_entry['Question']}

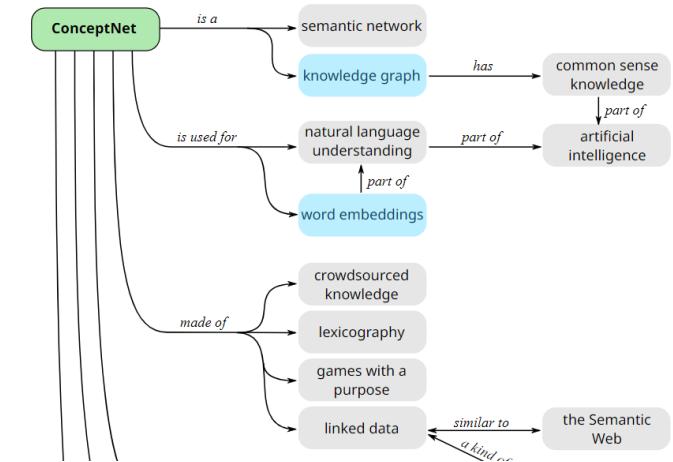
- A: {eval_entry["A"]}
- B: {eval_entry["B"]}
- C: {eval_entry["C"]}
- D: {eval_entry["D"]}

Keywords: {', '.join(final_keywords)}

Explanation:

Experiment: Task 2

- Task 2 - RoBERTa + Knowledge Graph + Graph Neural Network
 - Accuracy: 32.68%, 답변 생성 비율: 100.00%
 - 지식 그래프
 - 대규모 지식 그래프인 ConceptNet 기반
 - MMLU 데이터셋의 지식을 추가
 - 특수 토큰 추가: [1]
 - Fixed ConceptNet embedding
 - Trainable 5 layers GNN
 - Train-val split → Train data augmentation
 - Question의 단어 중 하나를 동의어로 대체 × 5개 = 6배의 Train data
 - Overfitting 방지를 위해 dropout rate = 0.5로 설정



Experiment: Task 2

- Task 2 - GPT-3.5, GPT-4
 - GPT-3.5 - Accuracy: 36.00%, 답변 생성 비율: 100.00%
 - GPT-4 - Accuracy: 65.37%, 답변 생성 비율: 100.00%
 - 아주 기본적인 prompt 사용, 한 번에 30개의 질문 입력
 - 성능 개선의 여지가 많음

Experiment: Task 2

- Task 2 - GPT-3.5, GPT-4



You

아래 질문들은 AI 또는 AI와 관련된 컴퓨터과학, 수학, 통계학 등의 지식에 대한 문제들이야. A, B, C, D 중에 적절한 답변을 고른 것을 나열해줘.

{"Question": "A local library has a scanner to detect library materials that have not been checked out. Each item has a chip somewhere inside. Upon checkout, the chip is deactivated so the scanner will not set off the alarm. The scanner has a 98% chance of detecting an active chip (meaning the material has not been checked out) and setting off the alarm. The scanner also has a 3% chance of sounding the alarm when someone passes through without an active chip. It is estimated that 0.5% of library customers actually try to leave the library with an active chip. What is the probability that, if the alarm sounds, the patron leaving the library has an item with an active chip?", "A": "0.0049", "B": "0.0348", "C": "0.141", "D": "0.97"}



ChatGPT

이 질문들에 대한 적절한 답변을 나열해보겠습니다.

1. 첫 번째 문제: 확률 계산

- 정답: **C** (0.141)

2. 두 번째 문제: 다항식의 완전제곱식 표현

- 정답: **D** ($-\frac{49}{12}$)

3. 세 번째 문제: 품질 관리와 가설 검정

- 정답: **C** (잘못된 사이즈의 볼 베어링 생산 계속)

4. 네 번째 문제: 정수의 공약수

- 정답: **A** ($\{-6, -2, -1, 1, 2, 6\}$)

5. 다섯 번째 문제: 확률과 독립 사건

- 정답: **D** (E와 F가 독립적이면서 상호 배타적일 수 없음)

Conclusion

- Task 1 & 2
 - 다양한 학습 방법을 통해 언어 모델이 데이터 사이언스 지식을 습득할 수 있었음
 - 학습 방법에 따라 지식 습득 및 추론의 결과가 매우 상이함

Conclusion

- Task 1: 직접 수집한 데이터셋으로 주관식 문제 평가
 - 대화를 통해 사용자는 데이터 사이언스 지식을 얻을 수 있음
 - Instruction tuning 이 가장 좋은 성능을 보여줌
- Task 2: MMLU 데이터셋으로 객관식 문제 평가
 - 객관식의 경우에도 역시 잘 정의된 1000개의 데이터만으로 효과적인 학습 가능
 - 소규모 LLM의 경우 fine-tuning과 prompt tuning을 함께하는 것도 좋은 방법
 - PLM을 잘 활용하는 것이 소규모 LLM보다 더 좋은 성능을 보임

Conclusion

- Future work
 - GPT-4 의 채점 결과를 믿을 수 있을까?
 - 길이가 긴 답변에 대해 더 높은 점수를 부여하는 경향이 있음
 - 더 정교하고 구체적인 Graph-of-thoughts 방법론 시도
 - Solar

